La educación es de todos
Mineducación

SABER AL DETALLE

EDICIÓN
08
Bogotá D.C.
Abril de 2021
ISSN: 2590 - 4663
Publicación Trimestral

Instituto Colombiano para la Evaluación de la Educación, ICFES
Oficinas: Calle 26 No. 69-76 . Torre 2, pisos 15 -18
Edificio Elemento, Bogotá . Colombia

Directora General: **Mónica Patricia Ospina Londoño**
Directora de Evaluación: **Natalia González Gómez**
Subdirectora de Análisis y Divulgación: **Mara Brigitte Bravo Osorio**

Subdirectora de Estadísticas: **Jeimy Paola Aristizábal Rodríguez**
Subdirección de Diseño de Instrumentos: **Javier Toro Baquero**
Coordinación General: **Dirección de Evaluación**

¿CUÁLES SON LOS MODELOS DE CALIFICACIÓN DE LAS PRUEBAS SABER?

icfes mejor saber

¿CUÁLES SON LOS MODELOS DE CALIFICACIÓN DE LAS PRUEBAS SABER?

El Instituto colombiano para la evaluación de la educación (Icfes) emplea diferentes modelos psicométricos para la medición de trazos latentes de los evaluados. En la mayoría de las pruebas, los trazos latentes corresponden a las habilidades y a los factores asociados al aprendizaje de los estudiantes. Cada uno de estos modelos tiene características propias de la medición: el tipo de escala de respuesta de los ítems (pregunta cerrada, es decir, de selección múltiple con única respuesta donde una respuesta correcta tiene 1 y una respuesta incorrecta tiene 0, pregunta abierta o escala tipo Likert) o el tamaño poblacional de los evaluados. En cualquier caso, los insumos empleados para el uso de los diferentes modelos de calificación son las respuestas de los evaluados a los ítems de la prueba de competencias que aplican los evaluados, de los cuestionarios auxiliares y la información socioeconómica. En esta edición, ahondaremos sobre los modelos psicométricos empleados para la calificación de los exámenes del Icfes y profundizaremos sobre sus características y diferencias.

¿Por qué se contemplan diferentes modelos psicométricos?

En Colombia, el Icfes aplica exámenes que buscan medir las diferentes habilidades, conocimientos y destrezas de los evaluados en distintos niveles educativos. Para ello, usa diferentes modelos de calificación que surgen del análisis psicométrico, la selección del modelo de calificación depende de diferentes factores. En el caso particular de los exámenes de Estado Saber Pro y Saber TyT, la selección del modelo depende de la clasificación del módulo. En el Icfes se tienen dos clasificaciones, módulos genéricos y módulos específicos[1], estos últimos a su vez se clasifican en dos, módulos adoptados y módulos no adoptados. Un módulo es "adoptado", si cumple con ciertos criterios de calidad en cuanto a la estructura interna y si estos se mantienen estables respecto a la confiabilidad de la prueba durante tres aplicaciones consecutivas, para estos módulos se utilizan los modelos de Teoría de Respuesta al Ítem (TRI). En contraste, un módulo que no cumple con las anteriores características es clasificado como "no adoptado", para estos módulos se emplea la Teoría Clásica del Test (TCT). Ahondaremos sobre amobos modelos más adelante. Además, dentro del análisis psicométrico de estos instrumentos de medición se tienen en cuenta características tales como, la consistencia interna (la confiabilidad sobre lo que se pretende medir), los tamaños poblacionales por aplicación y los parámetros de los ítems.

1. Esto aplica para el caso de los módulos específicos, ya que los genéricos se califican con un modelo específico establecido en la Resolución 268 de 2020.

¿Qué diferencias hay entre TCT y TRI?

La calificación aplicando procedimientos de TCT toma en consideración las respuestas correctas sobre el total de preguntas del examen, pues bajo este modelo la puntuación $\boldsymbol{X}$ obtenida en una prueba se puede expresar en términos de un puntaje verdadero $\boldsymbol{V}$ y un error de medición $\boldsymbol{e}$ que tiende a cero al evaluar una gran cantidad de veces el ítem (ver **Tabla 1**). El parámetro de dificultad, bajo este modelo, corresponde a la proporción de evaluados que responden correctamente un ítem. Una característica de la TCT es que no permite realizar comparaciones de los puntajes a través del tiempo, puesto que el parámetro de dificultad es sensible a la habilidad de la población evaluada. Teniendo en cuenta que el puntaje promedio de un evaluado corresponde a la razón entre el número de ítems correctos sobre el número de ítems de un examen, si se toman dos aplicaciones con poblaciones diferentes, aumentos en el puntaje promedio no necesariamente implican que los resultados estén mejorando a través del tiempo, sino a índices de dificultades diferentes para ambas poblaciones.

TABLA 1. Modelo TCT

Modelo	Función matemática
TCT	$X=V+e$

Fuente: Elaboración propia

De la primera publicación de Saber al detalle, recordemos que la TRI se utiliza para medir variables latentes, es decir, variables que no se pueden medir directamente a partir de una estimación de las respuestas dadas por los evaluados a un conjunto de preguntas. La TRI permite estimar la habilidad de los individuos (conocida como θ) con base en las respuestas dadas a los ítems y las características de estos, los cuales son estimados durante el proceso de calibración o a través de un proceso de anclaje[2]. De esta manera, es posible establecer una relación entre habilidad estimada de los evaluados y la probabilidad de acertar el ítem (teniendo en cuenta sus parámetros).

Bajo la TRI, la medida de dificultad está en la misma métrica que la habilidad y busca identificar la habilidad bajo el cual se espera que el 50% de los evaluados contesten correctamente el ítem, y por tanto ítems más fáciles reflejan índices de dificultad menores. A diferencia de la TCT, en la TRI existe un supuesto de invarianza de los parámetros de los ítems, es decir, que los parámetros del ítem no cambian, aunque los evaluados que contesten sean distintos, lo cual permite garantizar comparabilidad de resultados a través del tiempo. Como hemos mencionado en varias publicaciones, hay diferentes modelos de TRI que dependen del número de parámetros asociados a las características del ítem, además de la inclusión del parámetro de habilidad del

2. Para profundizar sobre los procedimientos de equiparación realizados en las pruebas Saber que permiten definir una métrica común entre aplicaciones refiérase a la publicación 3 de Saber al detalle: "¿Qué garantiza la comparabilidad de los resultados en las pruebas Saber realizadas por el Icfes?"

evaluado (ver **Tabla 2**). De hecho, la inferencia de que la habilidad influye sobre la probabilidad de responder correctamente a las preguntas, no existe bajo TCT, pues como se observa en la **Tabla 1** no hay relación entre el resultado del examen y la habilidad que mide, mientras que en TRI, la habilidad es un parámetro del modelo[3].

Cumpliendo los requerimientos de muestra mínima de individuos evaluados para tener estimaciones precisas de los parámetros de los ítems (de Ayala, 2009) es posible emplear la TRI como método de calificación con mayores bondades de análisis. Recordemos, por ejemplo, que una propiedad interesante de los modelos de TRI es que, la probabilidad de responder correctamente un ítem aumenta en la medida en que aumenta la habilidad del evaluado (la cual corresponde con la monotonicidad creciente de la función matemática implementada). Así mismo, cada evaluado cuenta con un índice de precisión del puntaje del examen a través del error estándar de medición.

Pese a las diferencias, los enfoques de la TCT y de la TRI son complementarios en el análisis psicométrico de los ítems de las pruebas del Icfes.

TABLA 2. Modelos TRI

Modelo	Número de parámetros	Función matemática
Modelo de dos parámetros logístico (2PL)	Dos parámetros (dificultad b_i y discriminación a_i) asociados con el ítem y habilidad estimada θ_j	$P(X_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$
Modelo de tres parámetros logístico (3PL)	Tres parámetros (dificultad b_i, discriminación a_i y pseudo-azar c_i) asociados con el ítem y habilidad estimada θ_j	$P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$

La función matemática corresponde a la probabilidad de que el evaluado j responda correctamente al ítem i, dada la habilidad y los parámetros asociados al ítem.

Fuente: Elaboración propia

3. Para profundizar sobre TRI, sus supuestos y las curvas características del ítem refiérase al documento 1 "¿Cómo se generan los puntajes en las pruebas Saber del Icfes?"

¿Por qué son complementarios los modelos TCT y TRI?

Los parámetros de dificultad (b_i), discriminación (a_i) y pseudo-azar (c_i) de los ítems y de habilidad (θ_j) de las ecuaciones de la **tabla 2** son desconocidos, pero conocemos las respuestas de los evaluados a las preguntas de cada una de las pruebas. Con base en ese insumo, se realiza un proceso que antecede al proceso de calificación que se conoce como análisis de ítem y permite comprender el comportamiento psicométrico de los ítems. En este proceso, se calculan algunas estadísticas descriptivas de los modelos bajo enfoques de TCT y de TRI, las cuales permiten crear alertas sobre el comportamiento de los ítems en términos de la omisión de respuesta (no respuesta), de distribuciones por tipo de respuesta, de dificultad, de correlaciones entre las preguntas yel puntaje del examen. Además la precisión del instrumento se mide a través de la evaluación por consistencia interna de la prueba la cual permite medir la magnitud en que los ítems de un test están correlacionados. Esta se mide a través del Alpha de Cronbach para ítems con respuestas polinómicas o mediante el estadístico de confiabilidad KR20 en el caso de ítems con respuestas dicotómicas (ver **Tabla 3**). Otros filtros adicionales de análisis de ítem corresponden a los relacionados con el análisis de copia.

TABLA 3. Estadísticos TCT y TRI para el análisis de los ítems

Modelo	Estadístico	Descripción
TCT	Tasa de omisión	Análisis de omisión (no respuesta) por ítem.
	Distribución de respuesta por opción de respuesta	Porcentaje de respuestas por opción de respuesta.
	Correlación biserial	Correlación entre las puntuaciones obtenidas por los individuos en el total de la prueba y las del ítem
	Alpha de Cronbach	Indicador de consistencia interna de la prueba.
	Índice de dificultad	Proporción de personas que contestaron correctamente el ítem
TRI	Unidimensionalidad	Factor dominante que dé cuenta de una variación considerable de respuestas.
	Alerta de parámetros	Se analiza la independencia de los parámetros de las personas y la de los parámetros de los ítems
	DIF	Análisis diferencial de los ítems por características de los individuos

Fuente: Icfes, 2020

Cabe señalar que, dado que el parámetro de habilidad θ_j y los parámetros de los ítems no son observables y la relación entre ellos no es lineal pues el modelo de calificación de TRI emplea una función logística, en el modelo se emplea un proceso de estimación a través de cuadraturas Gaussianas en el algoritmo de máxima verosimilitud marginal MML (Maximum Marginal Likelihood, por sus siglas en inglés) (Perozo, 2016) que permite encontrar los valores de los parámetros del modelo que maximizan la probabilidad de observar el conjunto de datos. Este proceso se realiza de forma iterativa hasta que la diferencia absoluta de las estimaciones de los parámetros en el paso i comparados con las estimaciones en el paso $i+1$ es menor o igual a un criterio de parada, que se denomina criterio de convergencia.

Como se observa, los modelos de calificación TCT y TRI son complementarios pues permiten generar análisis estadísticos con indicadores básicos de fácil interpretación y otros más robustos que permiten tener información que contribuya a un adecuado análisis de los ítems y las pruebas y la producción de resultados. Por lo cual, uno no es mejor que el otro, solo ofrecen información diferente.

¿Qué modelos psicométricos se emplean en la calificación de resultados?

En esa sección analizaremos los modelos de dos y tres parámetros que se emplean en la calificación de los exámenes de Estado Saber 11°, Saber Pro y Saber TyT. Recordemos que el examen Saber 11° está compuesto por 5 pruebas que corresponden a lectura crítica, matemáticas, sociales y ciudadanas, ciencias naturales e inglés. En el caso de los exámenes Saber Pro y Saber TyT, los evaluados presentan de manera obligatoria el componente de competencias genéricas compuesto por los módulos delectura crítica, razonamiento cuantitativo, competencias ciudadanas, comunicación escrita e inglés. Para algunos programas se ofertan módulos de competencias específicas que, según su Núcleo Básico de Conocimiento (NBC), indagan sobre competencias puntuales.

Un elemento particular tanto de la aplicación del examen Saber 11° como de los módulos de competencias genéricas de los exámenes Saber Pro y Saber TyT es que se mide al total de la población objetivo. Esto quiere decir que se evalúa a todos los estudiantes de educación media próximos a culminar grado 11 por calendario (en el caso de Saber 11°), así como a todos los estudiantes de carreras profesionales, técnicas y tecnológicas que hayan culminado al menos el 75% de su programa académico, independientemente del núcleo básico de conocimiento al que pertenezca (en el caso de Saber Pro y TyT). Al contar con el tamaño de muestra adecuado es posible emplear modelos de TRI, que, para el caso de 3PL se compone del parámetro asociado con la habilidad de los evaluados y tres parámetros asociados a características de los ítems: dificultad (b_i), discriminación (a_i) y pseudo-azar (c_i). Por su parte, un modelo 2PL se compone del parámetro asociado con la habilidad de los evaluados y dos parámetros asociados a características de los ítems: dificultad (b_i), discriminación (a_i). Es decir, para un modelo 2PL no se analiza el parámetro de pseudo-azar (c_i). Como se mencionó anteriormente la elección del

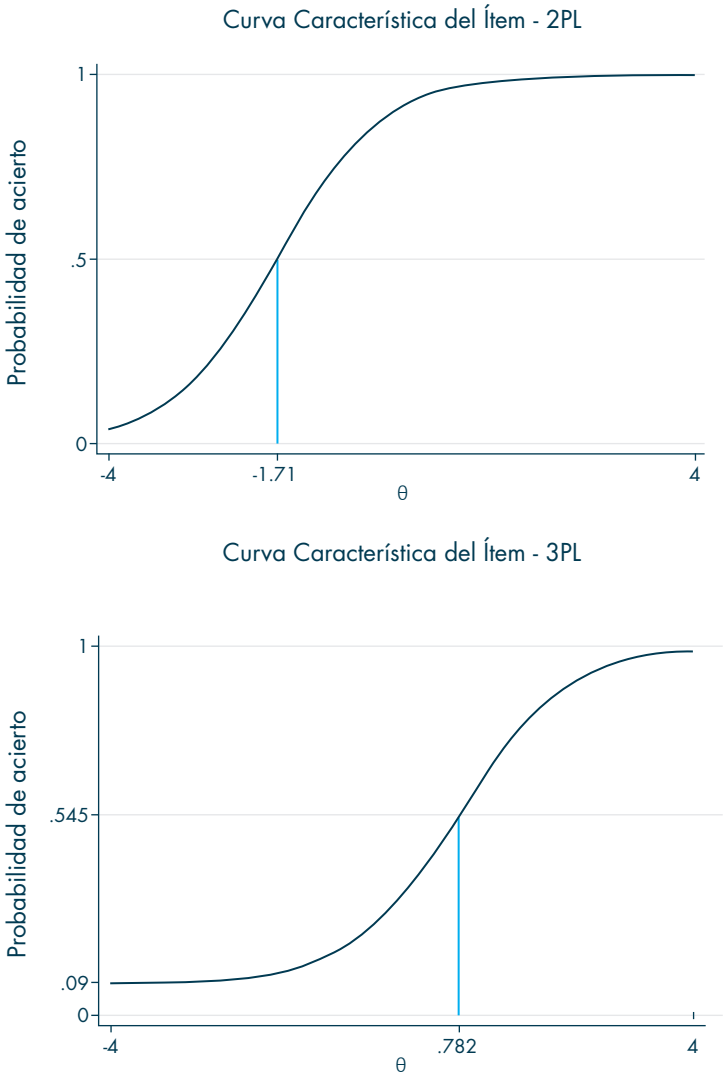
modelo depende del tipo de clasificación de los módulos en el caso de los exámenes Saber Pro y Saber TyT en los módulos específicos.

Un escenario distinto sucede para los módulos específicos, pues como su nombre lo indica, va dirigido a grupos poblacionales particulares según el NBC y por tanto, no se cuenta con las respuestas de todos los evaluados de los programas de educación superior para cada módulo. En tal caso, un modelo de 3 parámetros no resulta ser el más adecuado, pues cuando el tamaño de la muestra es pequeño, las estimaciones de los parámetros de los elementos tendrían errores estándar muy grandes y estarían sesgados si se usan para estimar las habilidades de los individuos (Linacre, 1994). Además, este tipo de módulos requiere un grado de estabilidad, precisión o consistencia como instrumentos de medición (Abad, Garrido, Olea, & Ponsoda, 2006). La mejor opción resulta ser el modelo de calificación TRI de 2 parámetros (2PL).

Dado que los dos modelos difieren en las características de los ítems que estiman, la información de dificultad y discriminación se observa para un modelo 2PL, mientras que en un modelo 3PL se reporta adicionalmente el pseudo azar. En la **figura 1** se observa la curva característica de un ítem de la prueba Saber TyT estimado bajo el modelo 2PL, y podemos observar que para una habilidad θ_j que para este caso puede tomar valores entre -4 (menor habilidad) y 4 (mayor habilidad) de -1.71 la probabilidad de acierto es del 50%, mientras que para un ítem diferente de Saber 11° cuya habilidad ha sido estimada bajo el modelo 3PL podemos analizar que con una probabilidad del 9% un estudiante con muy baja habilidad (-4) lograría contestar correctamente el ítem.

El modelo de calificación empleado para medir preguntas abiertas corresponde al modelo de múltiples facetas de Rasch (MMFR), que considera elementos tales como la severidad del evaluador (codificador), la dificultad de la tarea y la clasificación del evaluado que resultan relevantes para el proceso de calificación. Estos elementos, que se conocen como facetas son considerados para estimar la habilidad del evaluado. La metodología de este modelo la ahondamos a profundidad en la publicación anterior, por lo cual, los invitamos a referirse a dicha publicación para Saber al detalle ¿cómo se califican las preguntas de respuesta abierta en el contexto de las pruebas Saber?

FIGURA 1. Curvas características de los ítems según modelo TRI



Fuente: Icfes, 2020

Bibliografía

Abad, F. J., Garrido, J., Olea, J., & Ponsoda, V. (2006). *Introducción a la psicometría.* Madrid, España: Universidad Autónoma de Madrid.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*. Statistical theories of mental test scores.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Guilford Publications.

DeMars, C. E. (2010). *Item response theory.* Oxford University Press.

Perozo, M. F. (2016). *Una aplicación de valores plausibles a la calificación de pruebas estandarizadas vía simulación*. Comunicaciones en Estadística, 55-78.

SABER AL DETALLE

¿CUÁLES SON LOS MODELOS DE CALIFICACIÓN DE LAS PRUEBAS SABER?

@icfescol | ICFES | icfescol | YouTube: ICFES

Instituto Colombiano para la Evaluación de la Educación, ICFES
Oficinas: Calle 26 No. 69-76 . Torre 2, pisos 15 -18
Edificio Elemento, Bogotá . Colombia